DOCUMENT RESUME

ED 320 919 TM 014 984

AUTHOR	De Ayala, R. J.
TITLE	The Influence of Dimensionality on CAT Ability Estimation.
PUB DATE	Apr 90
NOTE	27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability Identification; *Adaptive Testing; Bayesian Statistics; *Computer Assisted Testing; Difficulty Level; *Estimation (Mathematics); Item Response Theory; Simulation; Testing Problems; *Test Items
IDENTIFIERS	*Item Dimensionality

ABSTRACT

The effect of dimensionality on an adaptive test's ability estimation was examined. Two-dimensional data sets, which differed from one another in the interdimensional ability association, the correlation among the difficulty parameters, and whether the item discriminations were or were not confounded with item difficulty, were generated for 1,600 simulated examinees. The generated data were used for Bayesian computerized adaptive testing (CAT) simulations (three-parameter logistic model), and the CAT ability estimates were compared with the simulated examinees' known abilities. The dimensionality of response data shifted the focus for the minimization of measurement errors from known abilities (with unidimensional data) to the average of the latent abilities (with bidimensional data). Three tables and 24 graphs summarize the study results. (Author/SLD)

The Influence of Dimensionality on CAT Ability Estimation

R.J. De Ayala

University of Maryland

Please send correspondence to :

R.J. De Ayala
Measurement, Statistics, and Evaluation
Benjamin Building
University of Maryland
College Park, MD 20742

Paper presented at the NCME Annual Meeting, April 1990, Boston

ABSTRACT

This study examined the effect of dimensionality on an adaptive test's ability estimation. Two-dimensional data sets were generated which differed from one another in the interdimensional ability association, the correlation among the difficulty parameters, and whether the item discriminations were or were not confounded with item difficulty. The generated data were used for Bayesian CAT simulations (three-parameter logistic model) and the CAT ability estimates were compared with the the simulees known abilities (θ_Ts). Results show that the dimensionality of the response data shifts the focus for the minimization of measurement errors from θ_T (with unidimensional data) to the average of the latent abilities (with bidimensional data).

Running Head : Dimensionality and CAT estimation

Key Words : CAT, dimensionality, IRT, Bayesian methods, computerized testing

Computerized adaptive testing (CAT) is concerned with the minimization of measurement errors in the estimation of an examinee's ability. To achieve this goal the examinee is administered items based on his or her current ability estimate. These items are selected such that the examinee is expected to have about a fifty percent chance of correctly answering the items. Some of CAT's benefits include equiprecise measurement throughout the ability continuum and adaptive tests which are shorter than the corresponding paper-and-pencil tests.

CATs typically are based on one of the dichotomous unidimensional IRT models, such as the three-parameter logistic (3PL) or Rasch models (e.g., McBride & Martin, 1983; Kingsbury & Houser, 1988). The development of the CAT item pool requires the identification of the data's dimensionality before fitting the IRT model. That is, although some items may be considered unidimensional, other test items may require more than one ability to obtain a correct response. For instance, correctly answering a mathematical word problem may be considered to be a function of reading and mathematical abilities. Implications of the violation of unidimensionality for CAT item pool development (e.g., equating, scale shrinkage) may be found in Doody-Bogan and Yen (1983) as well as in Yen (1985).

Multidimensional models have been developed in order to address the issue of multiple latent dimensions (e.g., McKinley & Reckase, 1983; Sympson, 1978). These models are classified as either compensatory or noncompensatory. Conceptually, a compensatory model is one in which an examinee's latent traits interact to produce a response to an item. This interaction may take the form of an examinee's facility in one latent trait (θ) compensating for a deficiency in another θ. In contrast, in a noncompensatory model the examinee's θs do not interact to yield a response. Although these models have been used in some research they have yet to obtain widespread acceptance or use in applications.

Given that the dimensionality assumption of unidimensional IRT subsumes the principle of local independence (Lord, 1980), violation of this assumption should affect the likelihood function used for parameter estimation. A number of studies (e.g., Ackerman, 1989; Way, Ansley, & Forsyth, 1988; Ansley & Forsyth, 1985; Reckase, 1979) have examined the effect of multidimensional response data on unidimensional IRT parameter estimates. These studies have been primarily concerned with the effects of dimensionality on the calibration of a multidimensional data set by either LOGIST (Wingerskey, Barton, & Lord, 1982) or BILOG (Mislevy & Bock, 1982). Although the models used for data generation differed, the results of these studies have found that dimensionality affects parameter estimation. In general, when a compensatory multidimensional IRT model was used for data generation $\hat{b}$ was found to be an estimate of the average of the true *b*s (Way et al., 1988), $\hat{a}$ was an estimate of the sum of a_1 and a_2 (Way et al., 1988), and ability estimates $\hat{\theta}$ to be an estimate of the average true θs (Ackerman, 1989; Way et al., 1988). In contrast, data generation using a noncompensatory model showed that $\hat{b}$ was an overestimate of or correlated more highly with b_1 than with b_2 (Ackerman, 1989; Way et al., 1988; Ansley & Forsyth, 1985), $\hat{a}$ was an estimate of the average of the true *a*s (Way et al., 1988; Ansley & Forsyth, 1985), and $\hat{\theta}$ to be an estimate of the average true θs (Ackerman, 1989; Way et al., 1988; Ansley & Forsyth, 1985). In general, these conclusions come from correlational analyses of the parameters with their estimates and an assessment of the accuracy of parameter estimation by the calculation of the mean absolute difference (a.k.a., MAD or AAD) across whichever was pertinent, examinees or items.

In general, studies which have investigated the operating characteristics of CAT have involved the simulation of unidimensional data and item pools (e.g., Weiss, 1982; McBride, 1977; Jensema, 1974). However, given "...that no actual psychological measurement instrument is likely to be exactly unidimensional..." the issue becomes one of whether the "...instrument is sufficiently unidimensional to allow application of IRT" (Hulin, Drasgow, & Parsons, 1983, p. 40). In live testings, where the possibility of less than ideal unidimensional data may exist, the primary concern has been with the estimation of the reliability and validity of CAT (e.g., McBride & Martin, 1983; Weiss & Kingsbury, 1984). Further, because in these studies the examinee's true ability is unknown the influence of dimensionality on the accuracy of ability parameter estimation cannot be investigated.

This study investigated the effect of varying degreee of dimensionality on CAT ability estimation. That is, an adaptive test based on unidimensional item parameter was administered to an simulee who used more than one ability to respond. Two-dimensional data sets were generated which differed from one another in the interdimensional ability association, the correlation among the difficulty parameters, and whether the item discriminations were or were not confounded with item difficulty. This latter factor is included because of Reckase, Carlson, Ackerman, and Spray's (1986) finding that upper deciles of a unidimensional ability differ mainly on θ_2 while at lower deciles the ability differed primarily on θ_1 (cited in Ackerman, 1989). Simulees with known abilities were administered unidimensional tests and their abilities estimated on the basis of their multidimensional responses. In contrast to the studies mentioned above (i.e., Ackerman, 1989; Way et al., 1988; Ansley & Forsyth, 1985), the accuracy and bias of the $\hat{\theta}$s at various points along the ability continuum was assessed.

METHOD

Data : The data were generated according to a multidimensional 3PL (M3PL) model (Doody-Bogan & Yen, 1983). This model requires a set of multidimensional θs as well as a set of (multidimensional) item parameters. The multidimensional θs were generated such that the examinee's ability on dimension 1 (θ_1) was evenly distributed between -3.0 and 3.0 using 0.4 logit interval between successive θ levels (i.e., for 100 examinees θ_1 =-3.0, for

100 examinees θ_1 =-2.6, etc.). The examinee's ability on the second dimension (θ_2) was derived from θ_1 by using Hoffman's (1959) technique for generating correlated data. For each of the 160C simulees θ_2 was obtained by randomly sampling a normal deviate (Z) from a unit normal curve and calculating :

$$\theta_2 = \theta_1 + (k / r)Z \qquad (1);$$

where $k=\sqrt{1-r^2}$, and r is the desired intercorrelation between θ_1 and θ_2. Four interdimensional θ correlations ($r_{\theta_1\theta_2}$) were investigated from extreme bidimensionality to almost unidimensionality; values for $r_{\theta_1\theta_2}$ were 0.03, 0.30, 0.60, 0.90.

In the following an item parameter's subscript refers to a dimension. The difficulty parameters (b_1 and b_2) were generated in a fashion analogous to the generation of θ_1 and θ_2. That is, the b_1 for sets of four items was fixed at every 0.1 logit between -3.5 and 3.5 (e.g., for 4 items b_1= -3.5, for 4 items b_1= -3.4, etc.). The b_2 for each of the 284 items was derived from the item's b_1 using the correlated generation method mentioned above. Three b_1b_2 correlations ($r_{b_1b_2}$) were used in the study, 0.03, 0.60, and 0.90.

The discrimination parameters (a_1 and a_2) were created by randomly sampling from a uniform distribution with a minimum value of 0.20 and a maximum value of 1.8. This set of *a*s was combined with the three sets of *b*s to form three item pools where all item pools had the same set of *a*s; this combination of the randomly ordered *a*s with the *b*s form form the nonconfounding condition. The confounding between *a*s and *b*s was obtained by sorting a_1 into ascending order and sorting a_2 into descending order (cf., Ackerman, 1989). The pseudo-guessing parameter, *c*, was set to 0.20.

The interdimensional correlations of 0.30, 0.60, and 0.90 were obtained from the literature (Ackerman,1989; Way et al.,1988; Ansley and Forsyth, 1985); the $r_{\theta_1\theta_2}$ = 0.03 was used as an approximation to $r_{\theta_1\theta_2}$ = 0.0 because this latter value could not be used with the Hoffman's technique. The $r_{b_1b_2}$ = 0.03 was obtained from Yen (1985), whereas the r_{b1b2} = 0.60 ($r_{b1b2}{}^2$=0.36) and r_{b1b2} = 0.90 ($r_{b1b2}{}^2$=0.81) were used to simulate moderate and high linear relationships. The minimum and maximum *a*s are the same as those in Ackerman (1989). The constant used ior *c* came from Way et al. (1988).

To summarize, the data generation was based on 6 different combinations of item parameters (3 levels of r_{b1b2} by 2 levels of confounding) and four levels of interdimensional ability association. The crossing of these three factors produced 24 response data sets. For each data set the true θ_Ts plus the relevant 284 true item parameters were used to generate binary response strings with a random error component for each simulated examinee. Generation of the binary response strings was accomplished by calculating for a given θ_T pair and a given item the probability of a correct response according to the M3PL model. To create the random error component for a response, a random number was selected from a uniform distribution [0,1] and compared to the calculated probability. If the random number was less than or equal to the calculated probability, then a response of 1 was produced (a correct answer), otherwise a 0 was generated (an incorrect response).

Program : A computer program was written that simulated a CAT based on the 3PL model and which used Bayesian ability estimation with Owens Bayes updating (i.e., Jensema's (1974) alpha technique) for item selection. The adaptive testing simulation was terminated when either of two criteria were met : a maximum of thirty items was reached or when a standard error of estimate (SEE) of 0.05 or less was obtained.

A unidimensional item pool was created for use with the Bayesian CAT. Discrimination, difficulty, and pseudo-guessing parameters were generated for 284 items. The discrimination (*a*) and pseudo-guessing (*c*) parameters were generated by random sampling from a uniform distribution with the following restrictions : (a) *a* were restricted to the inclusive range of 0.80..2.00; and (b) *c* were allowed to vary between 0.00 and 0.20. The difficulty parameters (*b*) were uniformly distributed between -3.5 and 3.5 (inclusive) with four items at each 0.10 of an interval (i.e., there were four items with *b*=

-3.5, four iterıs with b= -3.4, etc.). The use of multiple items at each 0.1 interval was done to ensure that items of appropriate difficulty would always be available for the Bayesian CAT's ability estimation. These item parameters values are consistent with desirable item pool characteristics (Patience and Reckase, 1980; Urry, 1977). Therefore, to each of the 1600 examinees in each of the 24- multidimensional response data sets a Bayesian CAT was administered.

Analyses : Analysis of the CAT simulations involved using root mean square error (RMSE), bias, and correlations (Pearson product-moment, Spearman rank-order) between the $\hat{\theta}$ and θ_1, θ_2, and between $\hat{\theta}$ and the average of θ_1 and θ_2 ($\bar{\theta}$). Descriptive statistics were calculated on the number of items administered, the $\hat{\theta}$s as well as on various item pool characteristics.

RESULTS

For the 0.03, 0.30, 0.60, and 0.90 interdimensional ability conditions the observed correlations were -0.028, 0.303, 0.590, and 0.964, respectively. Table 1 shows the item parameters' interdimensional correlations for the confounded and nonconfounded conditions. As can be seen, for the desired $r_{b_1b_2}$ of 0.03, 0.60, and 0.90 the observed correlations were 0.095, 0.678, and 0.946. In addition, for the confounded conditions the correlation between a_1 and b_1 approached -1.00 and between a_2 and b_1 approximated 1.00. The unidimensional item pool used for the CAT simulations had an average a of 1.410 (median of 1.421) and a mean c of 0.102 (median=0.101). The Pearson product-moment correlation between a and b for the unidimensional item pool was 0.077 (Spearman rank-order was 0.076).

Insert Table 1 about here

Table 2 shows the correlational analyses between $\hat{\theta}$ and θ_1, θ_2, and $\bar{\theta}$ for the nonconfounded conditions. As can be seen, for each level of the $r_{b_1b_2}$ factor the association between CAT $\hat{\theta}$ and ∂_1 and with θ_2 became increasingly stronger as $r_{\theta_1\theta_2}$

increased. The intercorrelation between *b*s appeared to have a slight effect on the correlation between $\hat{\theta}$ and θ_1 and between $\hat{\theta}$ and θ_2. Further, there was a slight decrease in the average number of items administered with increasing intercorrelation between the *b*s. Although for the $r_{b_1b_2} = 0.90$ and $r_{\theta_1\theta_2} = 0.90$ conditions there were minimal differences between $r_{\hat{\theta}\bar{\theta}}$, $r_{\hat{\theta}\theta_1}$, and $r_{\hat{\theta}\theta_2}$, for all combinations of the $r_{b_1b_2}$ and $r_{\theta_1\theta_2}$ factors the linear association between $\hat{\theta}$ and $\bar{\theta}$ was greater than for either $r_{\hat{\theta}\theta_1}$ or $r_{\hat{\theta}\theta_2}$.

Insert Table 2 about here

Figure 1 shows the RMSE analysis for the three levels of $r_{b_1b_2}$ and the $r_{\theta_1\theta_2} = 0.03$ and $r_{\theta_1\theta_2} = 0.90$ conditions; the differences in the plotted θ values reflect the differences in the $r_{\theta_1\theta_2}$ conditions. As can be seen, the RMSE with respect to $\bar{\theta}$ was less than that of the RMSE of either θ_1 or θ_2 for all nonconfounded conditions. In fact, the RMSE with respect to $\bar{\theta}$ for the $r_{\theta_1\theta_2} = 0.90$ condition is comparable to RMSE for when $r_{\theta_1\theta_2} = 0.03$, $r_{\theta_1\theta_2} = 0.30$, and $r_{\theta_1\theta_2} = 0.60$; the RMSE plots for these latter two conditions are the intermediate steps in the progression from $r_{\theta_1\theta_2} = 0.03$ RMSE plots to those of $r_{\theta_1\theta_2} = 0.90$. The RMSE with respect to $\bar{\theta}$ decreased slightly as $r_{b_1b_2}$ increased. As $r_{\theta_1\theta_2}$ increased, the RMSE of θ_1 or θ_2 approached that of $\bar{\theta}$.

Insert Figure 1 about here

As would be expected from a Bayesian CAT, the CAT overestimated low ability on θ_1 and θ_2 (i.e., $\theta_T < -2.0$) and underestimated high ability on θ_1 and θ_2 (i.e., $\theta_T > 2.0$); Figure 2 shows the bias plots for the nonconfounded conditions presented in Figure 1. As $r_{\theta_1\theta_2}$ increased the bias with respect to θ_1 and θ_2 decreased. For all combinations of the $r_{\theta_1\theta_2}$ and $r_{b_1b_2}$ factors, minimal bias was obtained when $\hat{\theta}$ was considered an estimate of $\bar{\theta}$. The $r_{b_1b_2}$ factor does not appear to have a meaningful effect on bias for θ_1, θ_2, and $\bar{\theta}$.

Insert Figure 2 about here

Table 3 presents the results from the confounded conditions. As was the case with the nonconfounded condition, $\hat{\theta}$ is more highly related to $\bar{\theta}$ than to either θ_1 or θ_2. In general, $r_{\hat{\theta}\theta_1}$ tends to be larger than $r_{\hat{\theta}\theta_2}$ for $r_{\theta_1\theta_2}$ values of 0.03 and 0.30, whereas for the $r_{\theta_1\theta_2} = 0.60$ and $r_{\theta_1\theta_2} = 0.90$ conditions the opposite is true. Unlike the nonconfounded condition, when $r_{\theta_1\theta_2} = 0.90$ the $r_{\hat{\theta}\theta_2}$ and $r_{\hat{\theta}\bar{\theta}}$ correlations are more similar to one one another and higher in magnitude than $r_{\hat{\theta}\theta_1}$. Further, for all combinations of the r_{b1b2} and $r_{\theta_1\theta_2}$ factors the average test length in the confounded condition was slightly less than the corresponding nonconfounded condition test length. The pattern of decreasing test length with increasing r_{b1b2} association was not as evident with the confounded condition as it was under the nonconfounded condition.

Insert Table 3 about here

Inspection of the confounded conditions' RMSE plots showed the same relationship between $\hat{\theta}$, $\bar{\theta}$, θ_1, and θ_2; Figure 3 contains the confounded condition sample RMSE plots for the same conditions presented in Figure 1. For the $r_{b1b2} = 0.03$ and $r_{b1b2} = 0.60$ conditions and for the approximate range $-2.0 \leq \theta \leq 2.0$, the RMSE for the confounded conditions are lower than those for the nonconfounded conditions, regardless of the $r_{\theta_1\theta_2}$ condition; as r_{b1b2} increases the difference in RMSEs diminishes. As was the case for the nonconfounded condition, the RMSE of θ_2 was less than that of θ_1 for high ability examinees for the $r_{b1b2} = 0.03$ condition. In contrast to the nonconfounded condition, the RMSE with respect to θ_1 was less for lower ability examinees than that of the RMSE θ_2. For all combinations of interdimensional ability and difficulty association the RMSE of $\bar{\theta}$ was less than that of θ_1 and θ_2.

Insert Figure 3 about here

Figure 4 presents the corresponding bias plots to those in Figure 2's, but for the confounded condition. As can be seen, compared to the nonconfounded condition there was less bias for θ_1 at low abilities, but no meaningful difference at upper abilities. Although at the r_{b1b2} = 0.03 and r_{b1b2} = 0.60 conditions there is no difference between the confounded and nonconfounded conditions in bias with respect to θ_2, for the r_{b1b2} = 0.90 there was an increase in bias for $\theta < -1.0$. For all interdimensional difficulty levels there was an increase in bias for θ_2 in the θ range 1.0 to 3.0. In general, as $r_{\theta_1\theta_2}$ increased this pattern was evident, although with decreasing levels of bias in the estimation of θ_1 and θ_2. As was the case with the nonconfounded condition, the bias in $\hat{\theta}$ with respect to $\bar{\theta}$ was less than that of estimating either θ_1 or θ_2, except when $r_{\theta_1\theta_2}$ = 0.90. In this latter condition, the differences in bias with respect to θ_1, θ_2 and $\bar{\theta}$, may not be considered meaningful by some; for this $r_{\theta_1\theta_2}$ condition there does not appear to be any difference in bias between the confounded and nonconfounded conditions. For $r_{\theta_1\theta_2}$ = 0.90 and regardless of r_{b1b2} level, the CAT overestimated low ability more than it underestimated high ability.

Insert Figure 4 about here

As stated above, for the nonconfounded condition there was a slight decrease in the average number of items administered with increasing r_{b1b2}, although this pattern was not as evident with the confounded condition. Calculation of the average number of items administered at each of the 16 levels of θ showed that, in general, shorter tests were administered for $\theta < 0.0$ (e.g., average test lengths of 15-16 items depending on the condition) to longer tests for $\theta > 2.0$ (e.g., mean test lengths of 17-20 items depending on particular data set; the r_{b1b2} = 0.03, $r_{\theta_1\theta_2}$ = 0.90 condition had an atypical mean test length of 22 items for θ = 3.0). With increasing r_{b1b2} and $r_{\theta_1\theta_2}$ the mean test lengths became less variable across θ. Of the 38,400 adaptive tests simulated the absolute maximum and minimum test lengths were 28 and 11 items, respectively.

Conclusion and Discussion

In general, increasing interdimensional difficulty association produced a slight decrease on test length and an increase in the accuracy of ability estimation as assessed by RMSE. The associations between $\hat{\theta}$ and $\bar{\theta}$, θ_1, and θ_2 increased as the correlation between interdimensional difficulties and interdimensional ability increased. The largest associations were between $\hat{\theta}$ and $\bar{\theta}$; 0.957 and 0.961 for the nonconfounded and confounded conditions, respectively. For comparative purposes, a Bayesian CAT (maximum test length of 20 items and termination SEE of 0.05) using a unidimensional data set (generated according to the 3PL model and using this item pool) had a $r_{\hat{\theta}\theta}$ of 0.988 (for both Pearson and Spearman coefficients) and an average test length of 15.613.

When discrimination was confounded with difficulty, the ability estimates showed a differential association with one of the two latent traits, however, the correlation between $\hat{\theta}$ and $\bar{\theta}$ was always greater than that of $r_{\hat{\theta}\theta_1}$ and $r_{\hat{\theta}\theta_2}$. For all combinations of the r_{b1b2} and $r_{\theta_1\theta_2}$ factors the correlation between $\hat{\theta}$ and $\bar{\theta}$ for the confounded condition was always greater than for the correlation for the corresponding nonconfounded condition.

From the results of the studies on the effects of dimensionality on the calibration of compensatory multidimensional data it may be hypothesized that the finding that $\hat{\theta}$ was an estimate of the average true θs was, in part, a result of the fact that $\hat{b}$ was an estimate of the average of the true *b*s. That is, because *b* and θ are on the same scale, when the separate dimensions are collapsed in the estimation of *b*, the subsequent stage of estimating θ will also reflect the collapsed difficulty scale; both BILOG and LOGIST obtain $\hat{b}$s prior to estimating θ. However, given that in CAT the item parameters are assumed true then the collapsing of the two difficulty scales does not account for $\hat{\theta}$ being an estimate of the average θ_Ts.

Conceptually, the item pool may be considered to have come from the calibration of a unidimensional data set. However, the results should be generalizable to those situations where item parameters are obtained from data which are not truly unidimensional (i.e., the situations investigated by Ackerman, 1989; Way, et al, 1988; Ansley & Forsyth, 1985). For item selection it is the distribution of *b* and the magnitude of *a* and *c* which are important; CAT makes no distinction with respect to whether $\hat{b} = b$ or $\hat{b} = (b_1 + b_2)/2$ and $\hat{a} = a$ or $\hat{a} = a_1 + a_2$.

As stated above, CAT is concerned with minimizing the measurement errors associated with the estimation of an examinee's ability. It was shown that the dimensionality of the response data shifts the focus for the minimization of measurement errors from θ_T (with unidimensional data) to the average of the latent abilities (with bidimensional data). Although the results may be considered problematic by some, there may be situations where one is only interested in ordering examinees on their ability to perform or solve certain types of problems and not in ordering them on the separate latent abilities which may be required to solve the problems. For example, on a statistics exam the instructor may only be interested in a student's understanding of the appropriateness and use of t-tests. The problems may be stated as word problems and require stating the appropriate statistical hypotheses, identification of and calculating the relevant t-statistic, arriving at conclusions concerning the truth or falsity of hypotheses, etc. Most likely the instructor is not interested in the student's standing on the separate abilities required to answer the problem (e.g., his or her reading ability, math ability, etc), but in the student's understanding of t-tests. Reckase, Ackerman, and Carlson (1988) have concluded that IRT's unidimensionality assumption does not necessarily require test items to measure a single ability, but rather the unidimensionality assumption requires the test items to measure the same composite of abilities. For this study, this composite of abilities was the average of θ_1 and θ_2.

REFERENCES

Ackerman, T.A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.

Ansley, T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Doody-Bogan, E. & Yen, W.M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model.* Paper presented at the annual meeting of American Educational Research Association, Montreal.

Hoffman, P.J. (1959). Generating variables with arbitrary properties. *Psychometrika, 24*, 265-267.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory : Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Jensema, C.J. (1974). The validity of Bayesian tailored testing. *Educational and Psychological Measurement, 34*, 757-766.

Kingsbury, G.G. & Houser, R.L. (1988, April). *A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing.* Paper presented at the annual meeting of American Educational Research Association, New Orleans.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

McBride, J.R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (ed.), *New Horizons in Testing* (pp 223-237). New York : Academic.

McBride, J.R. (1977). Some properties of a Bayesian adaptive testing strategy. *Applied Psychological Measurement, 1*, 121-140.

McKinley, R. & Reckase, M.D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR83-2). Iowa City, IA : American College Testing Program.

Mislevy, R.J. & Bock, R.D. (1982). *BILOG, maximum likelihood item analysis and test scoring: Logistic model.* Mooresville, IN: Scientific Software, Inc.

Patience, W.M. & Reckase, M.D. (1980). *Effects of program parameters and item pool characteristics on the bias of a three-parameter tailored testing procedure.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Reckase, M.D., Ackerman, T.A., & Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193-203.

Reckase, M.D., Carlson, J.E., Ackerman, T.A., & Spray, J.A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the annual meeting of Psychometric Society, Toronto.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests : results and implications. *Journal of Educational Statistics, 4*, 207-230.

Sympson, J.B. (1978). A model for testing with multidimensional items. In D.J. Weiss (ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp 82-98). Minneapolis: University of Minnesota, Psychometric Methods Program, Department of Psychology.

Urry, V.W. (1977). Tailored testing : a successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Way, W.D., Ansley, T.N., & Forsyth, R.A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239-252.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Service.

Yen,W.M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*, 399-410.

Table 1. Item parameters interdimensional correlations[a].

Condition	Item	Parameter	a_2	b_1	b_2
$r_{b_1b_2} = 0.03$		a_1	-0.990	-0.995	-0.088
			(-0.032)	(-0.022)	(0.055)
		a_2		0.997	0.112
				(0.014)	(-0.034)
		b_1			0.095
					(0.095)
$r_{b_1b_2} = 0.60$		a_1	-0.990	-0.997	-0.677
			(-0.032)	(-0.022)	(0.027)
		a_2		0.995	0.671
				(0.014)	(-0.016)
		b_1			0.678
					(0.678)
$r_{b_1b_2} = 0.90$		a_1	-0.990	-0.997	-0.944
			(-0.032)	(-0.022)	(-0.002)
		a_2		0.995	0.940
				(0.014)	(0.002)
		b_1			0.946
					(0.946)

[a] Pearson product-moment correlations for confounded and (nonconfounded) conditions.

Table 2. Intercorrelations[a] between $\hat{\theta}$ and $\theta_1, \theta_2, \bar{\theta}$ and the average number the of items administered (Mean NIA) for the nonconfounded conditions.

Item Pool Characteristics		$r_{\hat{\theta}\theta_1}$	$r_{\hat{\theta}\theta_2}$	$r_{\hat{\theta}\bar{\theta}}$	Mean NIA (SD NIA)
$r_{b_1 b_2}$	$r_{\theta_1\theta_2}$				
0.03	0.03	0.500	0.645	0.821	17.206
		(0.518)	(0.520)	(0.785)	(2.577)
	0.30	0.611	0.752	0.844	17.187
		(0.630)	(0.660)	(0.825)	(2.649)
	0.60	0.741	0.816	0.873	17.146
		(0.751)	(0.825)	(0.854)	(2.587)
	0.90	0.890	0.893	0.900	17.408
		(0.890)	(0.880)	(0.891)	(2.773)
0.60	0.03	0.513	0.694	0.866	16.896
		(0.539)	(0.564)	(0.840)	(2.390)
	0.30	0.678	0.779	0.903	16.878
		(0.697)	(0.686)	(0.888)	(2.440)
	0.60	0.799	0.849	0.924	16.703
		(0.801)	(0.863)	(0.902)	(2.442)
	0.90	0.922	0.929	0.934	16.926
		(0.921)	(0.915)	(0.924)	(2.543)
0.90	0.03	0.552	0.723	0.914	16.407
		(0.562)	(0.601)	(0.889)	(2.249)
	0.30	0.727	0.794	0.942	16.323
		(0.734)	(0.707)	(0.925)	(2.243)
	0.60	0.829	0.874	0.955	16.265
		(0.823)	(0.886)	(0.928)	(2.257)
	0.90	0.945	0.951	0.957	16.428
		(0.940)	(0.935)	(0.942)	(2.408)

[a] Pearson product-moment correlation coefficient (Spearman rank-order correlation coefficient)

Table 3. Intercorrelations[a] between $\hat{\theta}$ and $\theta_1, \theta_2, \bar{\theta}$ and the average number the of items administered (Mean NIA) for the confounded conditions.

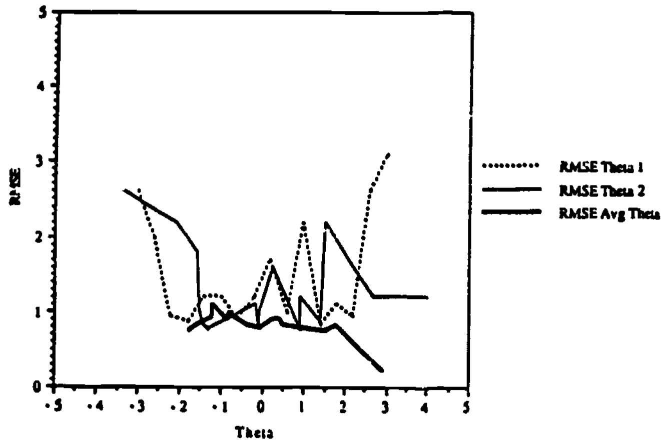
Item Pool Characteristics		$r_{\hat{\theta}\theta_1}$	$r_{\hat{\theta}\theta_2}$	$r_{\hat{\theta}\bar{\theta}}$	Mean NIA (SD NIA)
$r_{b_1 b_2}$	$r_{\theta_1 \theta_2}$				
0.03	0.03	0.628	0.623	0.897	16.433
		(0.674)	(0.437)	(0.851)	(2.213)
	0.30	0.692	0.768	0.905	16.689
		(0.747)	(0.617)	(0.886)	(2.379)
	0.60	0.763	0.850	0.905	16.703
		(0.801)	(0.836)	(0.883)	(2.495)
	0.90	0.907	0.921	0.922	17.394
		(0.907)	(0.910)	(0.909)	(3.178)
0.60	0.03	0.663	0.620	0.920	16.337
		(0.720)	(0.422)	(0.877)	(2.300)
	0.30	0.760	0.750	0.936	16.293
		(0.797)	(0.614)	(0.923)	(2.209)
	0.60	0.834	0.856	0.948	16.276
		(0.850)	(0.851)	(0.916)	(2.214)
	0.90	0.928	0.942	0.943	16.730
		(0.927)	(0.933)	(0.930)	(2.582)
0.90	0.03	0.704	0.608	0.941	15.917
		(0.749)	(0.418)	(0.907)	(2.110)
	0.30	0.787	0.756	0.956	16.066
		(0.802)	(0.635)	(0.942)	(2.085)
	0.60	0.840	0.874	0.961	16.031
		(0.845)	(0.868)	(0.923)	(2.056)
	0.90	0.944	0.955	0.958	16.230
		(0.939)	(0.945)	(0.941)	(2.235)

[a] Pearson product-moment correlation coefficient (Spearman rank-order correlation coefficient)

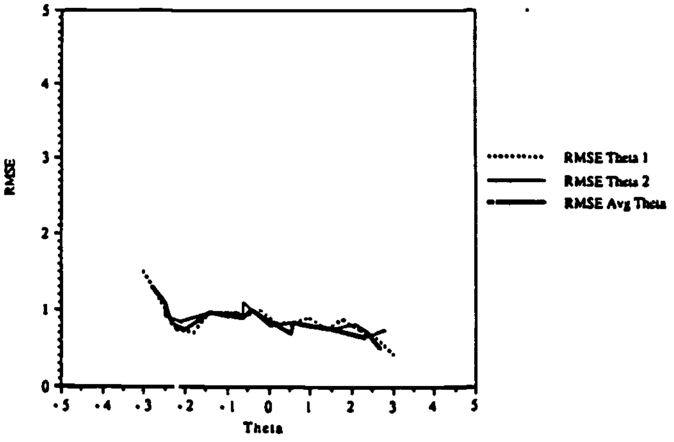
Figure Captions

Figure 1. RMSE analysis for the nonconfounded condtions for $r_{b_1b_2}$ = 0.03, $r_{b_1b_2}$ = 0.60, $r_{b_1b_2}$ = 0.90 and $r_{\theta_1\theta_2}$ = 0.03, $r_{\theta_1\theta_2}$ = 0.90.

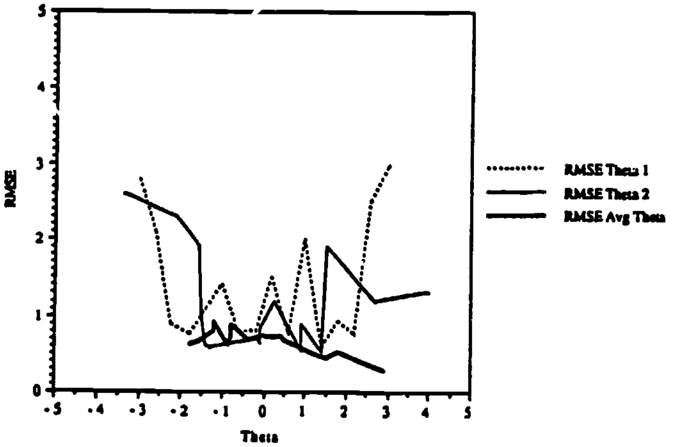
RMSE
Ability r=0.03; Difficulty r=0.03; Noncounfounded



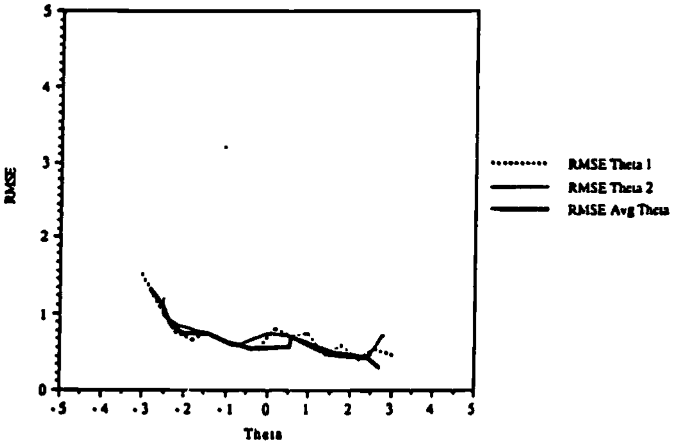
RMSE
Ability r=0.90; Difficulty r=0.03; Nonconfounded



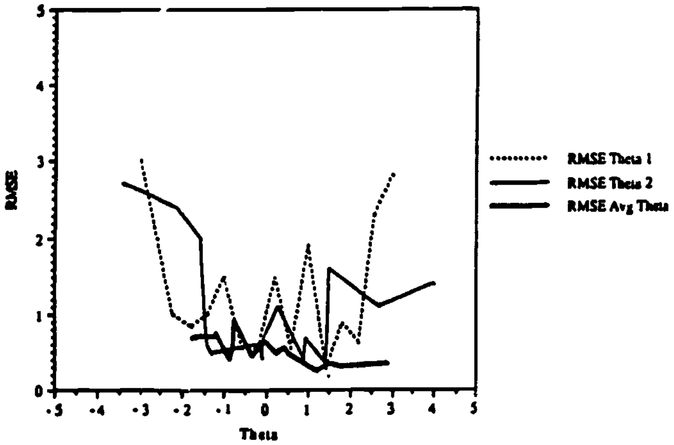
RMSE
Ability r=0.03; Difficulty r=0.60; Nonconfounded



RMSE
Ability r=0.90; Difficulty r=0.60; Nonconfounded



RMSE
Ability r=0.03; Difficulty r=0.90; Nonconfounded



RMSE
Ability r=0.90; Difficulty r=0.90; Nonconfounded

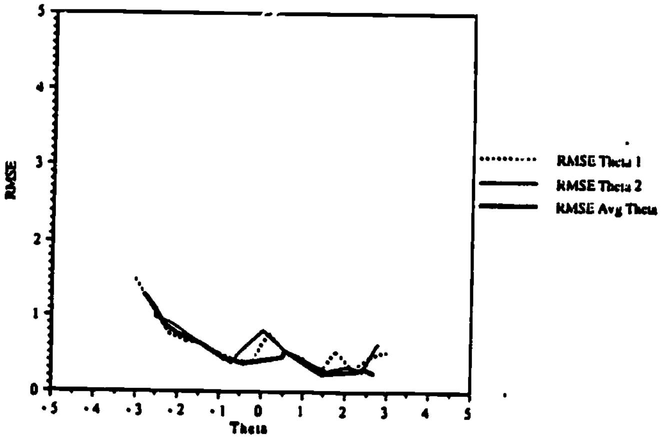
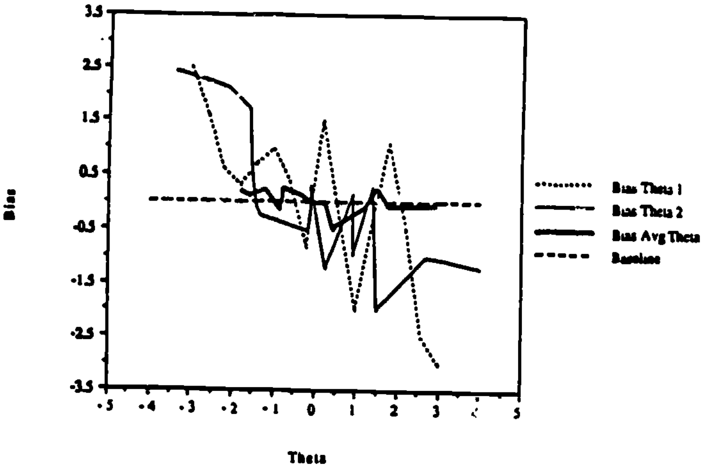


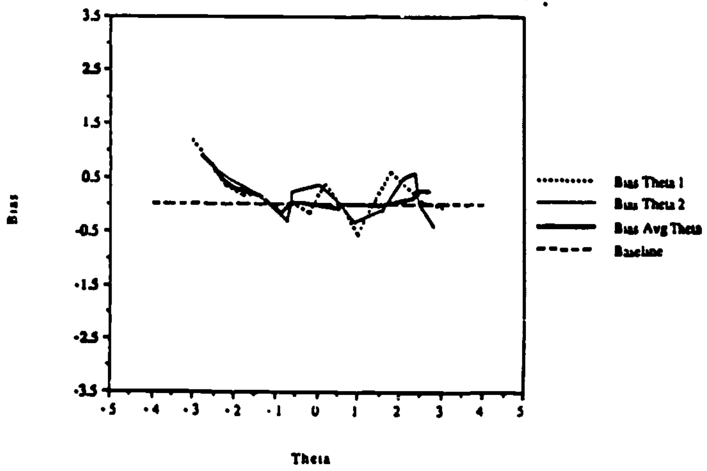
Figure Captions

Figure 2. Bias analysis for the nonconfounded condtions for $r_{b_1b_2} = 0.03$, $r_{b_1b_2} = 0.60$, $r_{b_1b_2} = 0.90$ and $r_{\theta_1\theta_2} = 0.03$, $r_{\theta_1\theta_2} = 0.90$.

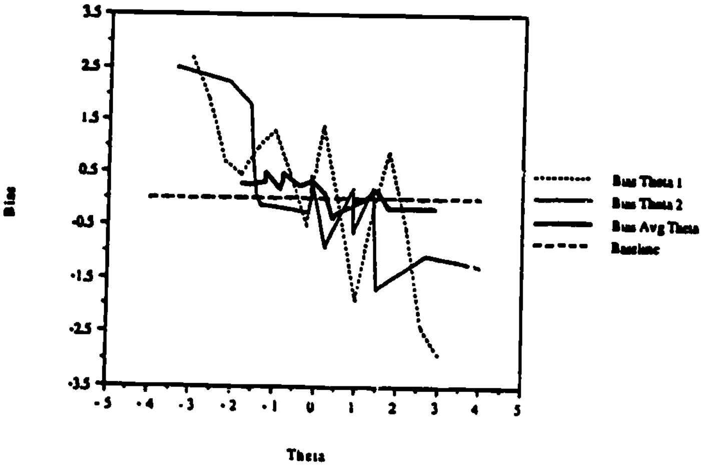
Bias
Ability r=0.03; Difficulty r=0.03; Nonconfounded



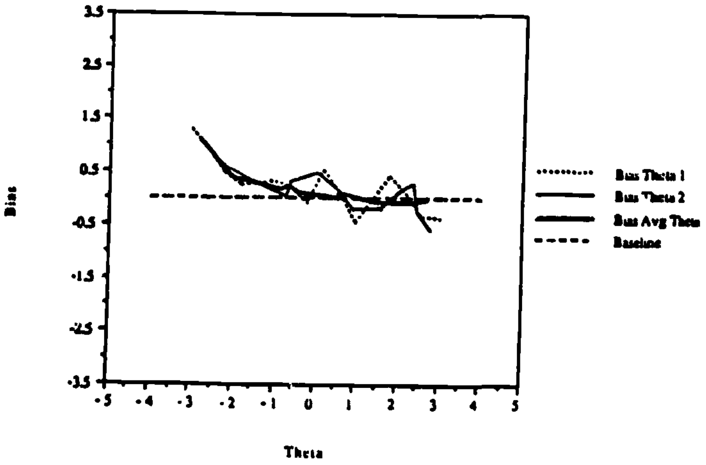
Bias
Ability r=0.90; Difficulty r=0.03; Nonconfounded



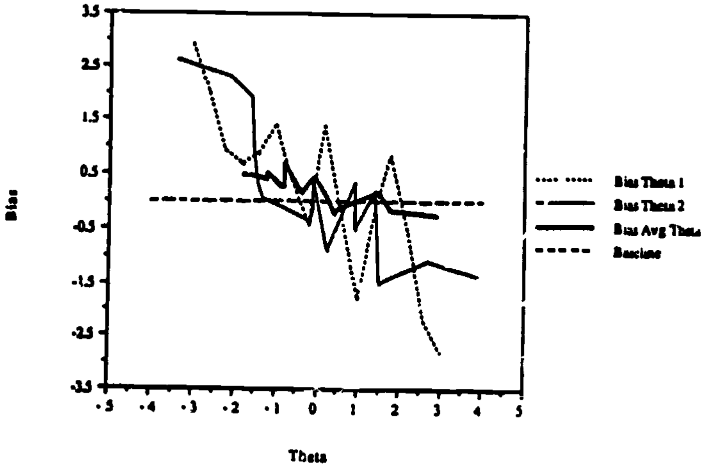
Bias
Ability r=0.03; Difficulty r=0.60; Nonconfounded



Bias
Ability r=0.90; Difficulty r=0.60; Nonconfounded



Bias
Ability r=0.03; Difficulty r=0.90; Nonconfounded



Bias
Ability r=0.90; Difficulty r=0.90; Nonconfounded

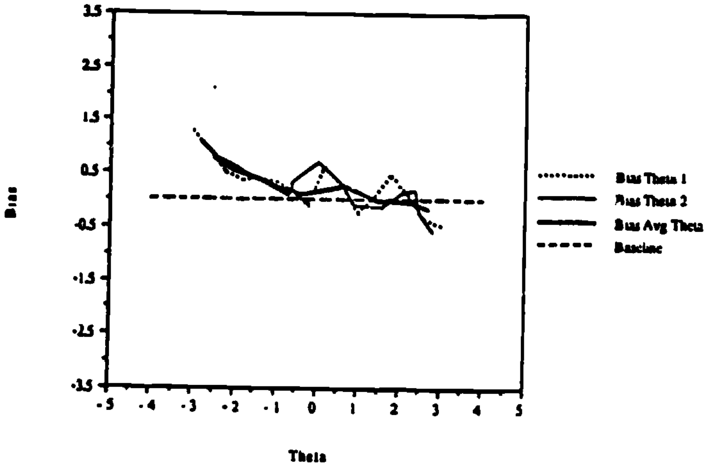
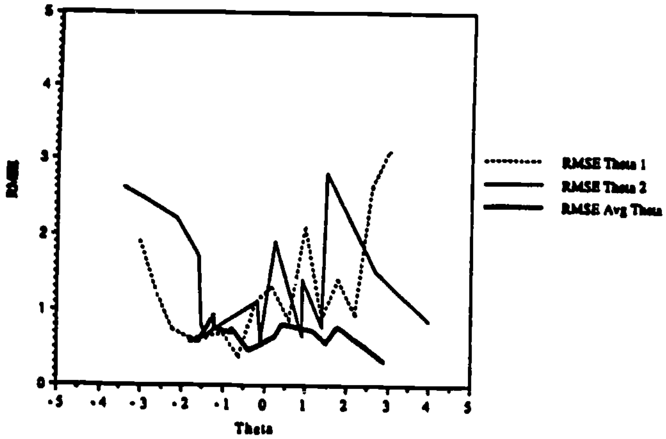


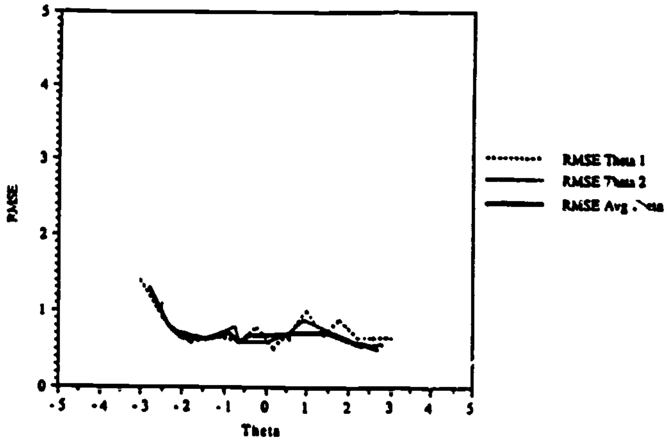
Figure Captions

Figure 3. RMSE analysis for the confounded condtions for $r_{b_1 b_2}$ = 0.03, $r_{b_1 b_2}$ = 0.60, $r_{b_1 b_2}$ = 0.90 and $r_{\theta_1 \theta_2}$ = 0.03, $r_{\theta_1 \theta_2}$ = 0.90.

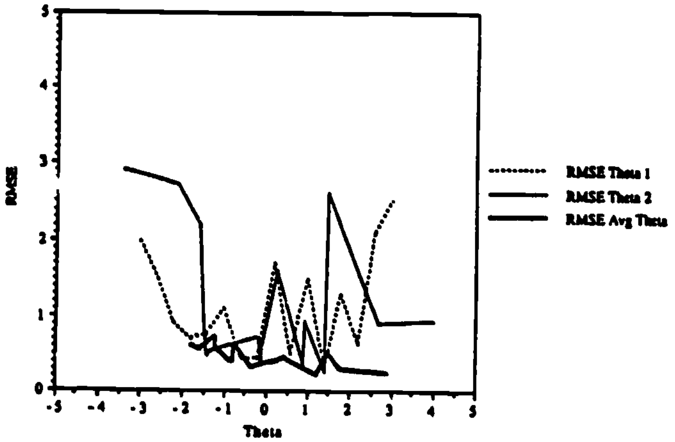
RMSE
Ability r=0.03; Difficulty r=0.03; Confounded



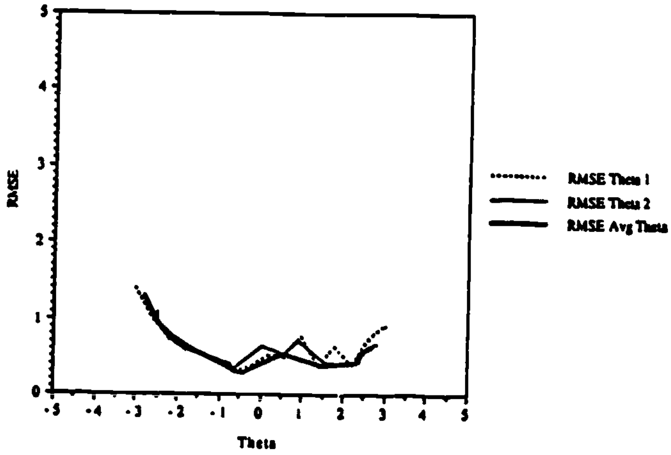
RMSE
Ability r=0.90; Difficulty r=0.03; Confounded



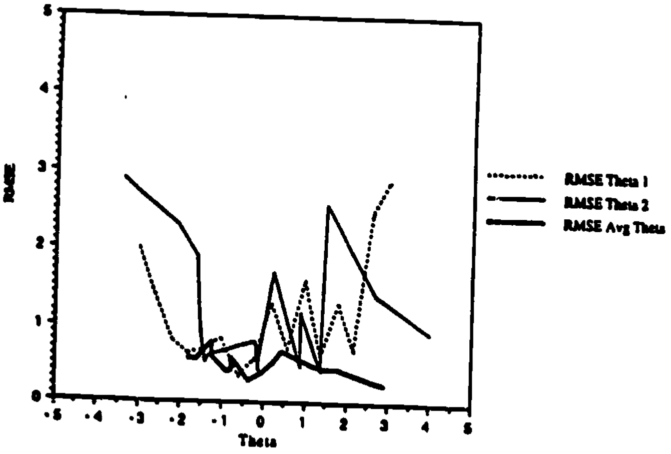
RMSE
Ability r=0.03; Difficulty r=0.90; Confounded



RMSE
Ability r=0.90; Difficulty r=0.60; Confounded



RMSE
Ability r=0.03; Difficulty r=0.60; Confounded



RMSE
Ability r=0.90; Difficulty r=0.90; Confounded

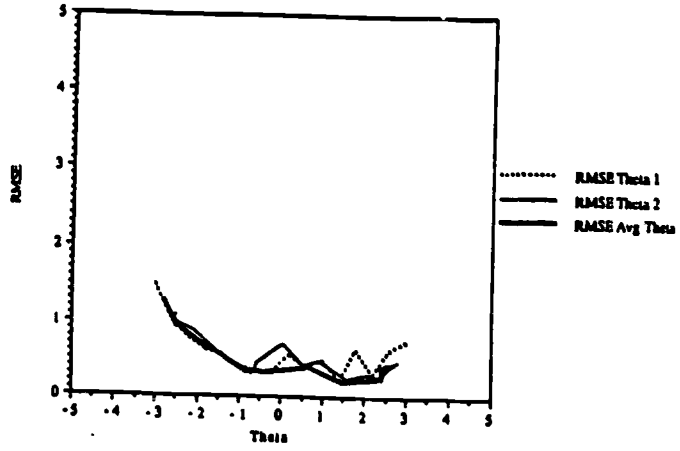
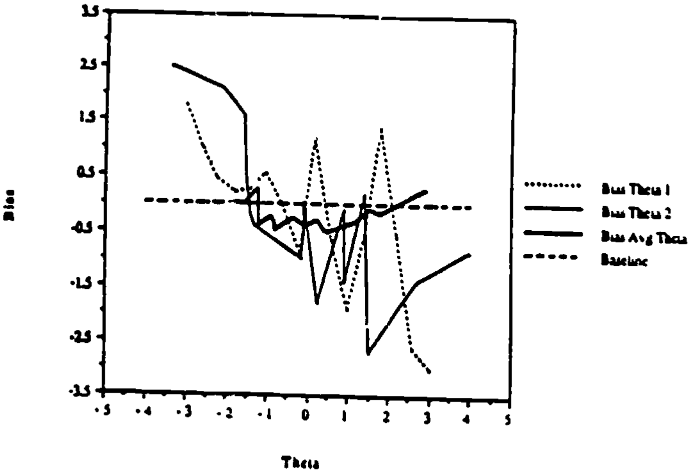


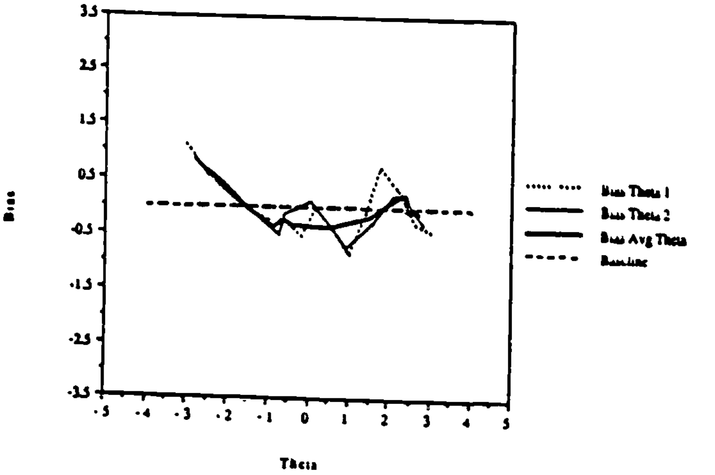
Figure Captions

Figure 4. Bias analysis for the confounded condtions for $r_{b_1b_2} = 0.03$, $r_{b_1b_2} = 0.60$, $r_{b_1b_2} = 0.90$ and $r_{\theta_1\theta_2} = 0.03$, $r_{\theta_1\theta_2} = 0.90$.

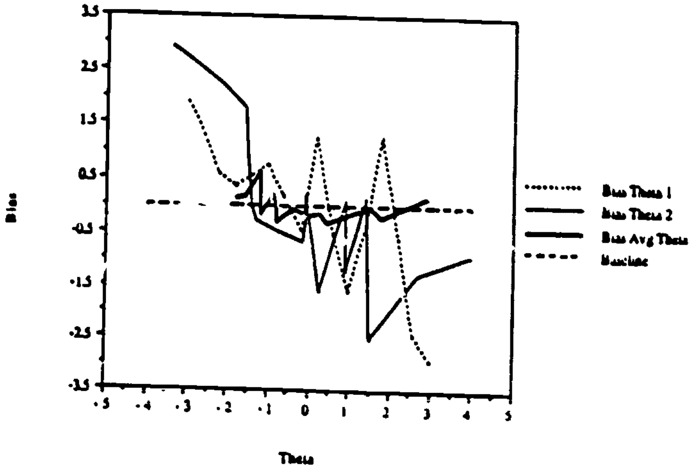
Bias
Ability r=0.C3; Difficulty r=0.03; Confounded



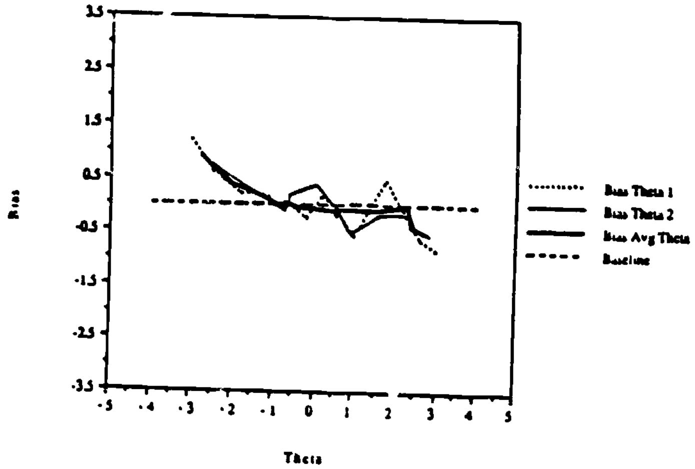
Bias
Ability r=0.90. Difficulty r=0.03; Confounded



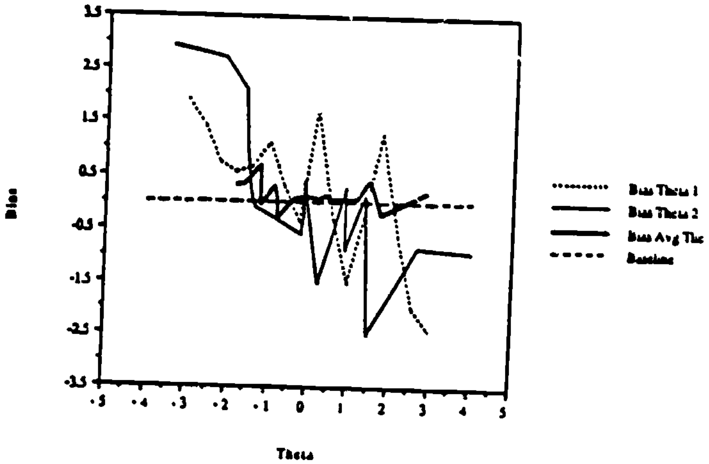
Bias
Ability r=0.03; Difficulty r=0.60, Confounded



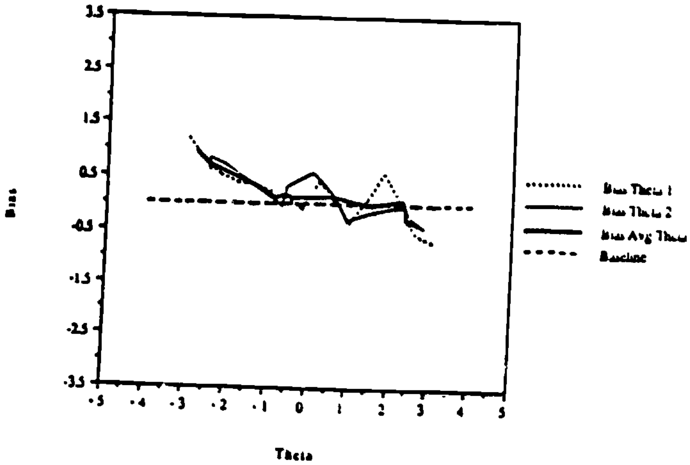
Bias
Ability r=0.90, Difficulty r=0.60; Confounded



Bias
Ability r=0.03, Difficulty r=0.90; Confounded



Bias
Ability r=0 90; Difficulty r=0 90, Confounded


Bias
Theta
Bias Theta 1
Bias Theta 2
Bias Avg Theta
Baseline